Supplementary Text

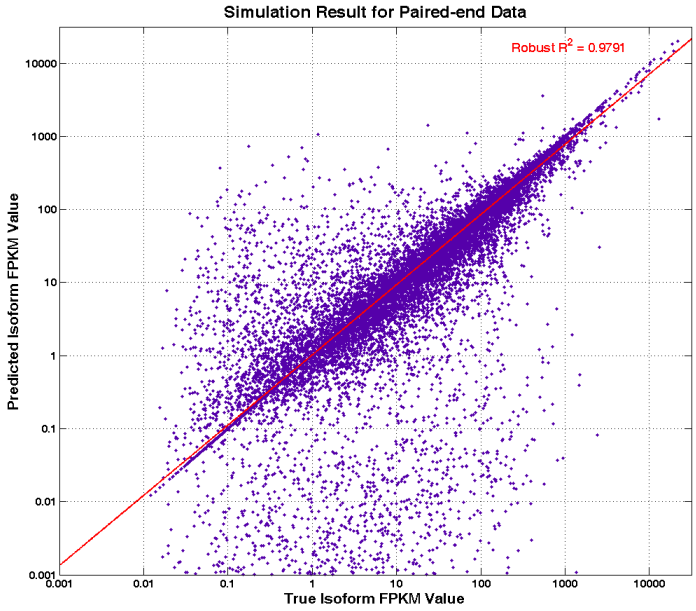

Figure S1. The simulation study to evaluate the accuracy of isoform quantification using the FluxSimulator. For paired-end short read data, we plot predicted isoform abundance scores against true abundance scores. R^2 calculated by robust linear regression analysis was shown in this figure. 30 millions 50-mer paired-end short reads were generated and used for this simulation study.

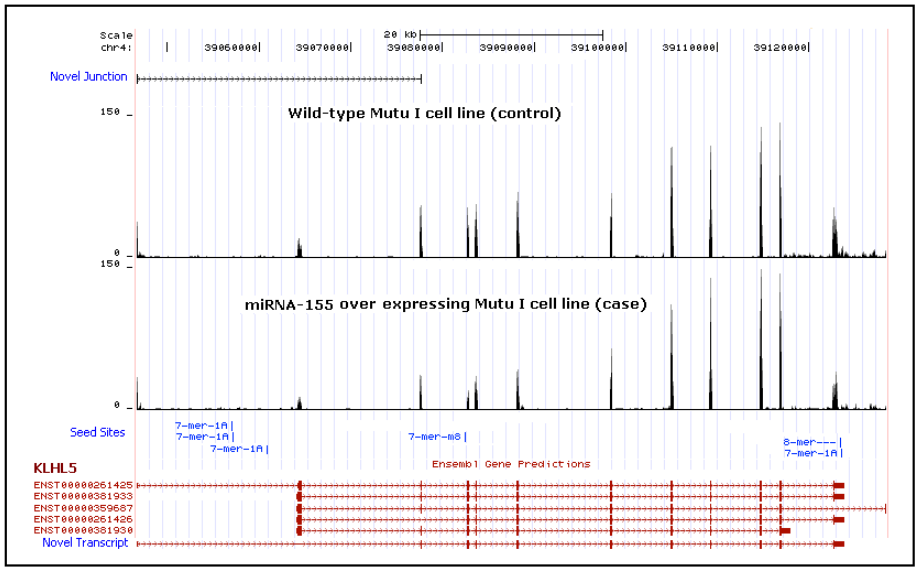

Figure S2. An example of novel (context specific) isoform target.

Section 1: Technical details of the EM algorithm for isoform quantification

Suppose for each gene there are J annotated isoforms (denoted as $I_1, I_2, \dots, I_j$). For each short read we observed, p_j is used to denote the probability that this short read is generated from isoform I_j, where $j = 1, \dots, J$ and $p_1 + p_2 + \dots + p_j = 1$.

Suppose for one gene, we have N short reads (denoted as $R_1, R_2, \dots, R_N$) and we know the correspondence between short reads and isoforms. Then we can use an $N \times J$ indicator matrix $Z = (z_{ij})_{i=1,\dots,N, j=1,\dots,J}$ to represent the correspondence between short reads and isoforms. If the ith read is generated from isoform I_j, then $z_{ij} = 1$; $z_{ij} = 0$ otherwise. So, for the Z matrix, each row indicates one short read, and only one column for this row with value equal to 1. If the Z matrix is our observed matrix, it is easy for us to calculate the isoform proportions. The probabilities $(p_j, j = 1, \dots, J)$ can be used for isoform proportion estimation. Intuitively, the number of short read which are generated from isoform I_j can be calculated by sum of jth column, correspondingly $n_j = \sum_{i=1}^{N} z_{ij}$, then the estimated isoform proportion $p_j = n_j/N$.

The ambiguity issue in our study is that most of short reads are compatible with more than one isoform, therefore, the indicator matrix Z is not fully observed. What we actually observed in a RNA-seq experiment is another matrix $Y = (y_{ij})_{i=1,\dots,N, j=1,\dots,J}$, where $y_{ij} = \frac{1}{l_j}$ if the ith short read is compatible with isoform I_j, and $y_{ij} = 0$ otherwise. l_j is the length of isoform I_j and $\frac{1}{l_j}$ measures the contribution of one short read to isoform I_j. Compare with matrix Z, which has one and only one non-zero value in each row, the matrix Y has one or more than one non-zero value in each row. If $y_{ij} = 0$, then z_{ij} must be 0, but if $y_{ij} \neq 0$, then z_{ij} may or may not be 1. In our manuscript, we define Y as the observed cDNA fragment-compatible matrix, and Z as the unobserved cDNA fragment-originating matrix.

Let's denote $P = (p_1, p_2, \dots, p_j)$. The log likelihood function of our isoform proportion model with the observed cDNA fragment-compatible matrix is

$$L(P \mid Y) = \sum_{i=1}^{N} \log\left(\sum_{j=1}^{J} y_{ij} \times p_j\right).$$

The maximum likelihood estimates (MLE) of P can be written as $\arg\max_P L(P \mid Y)$, which cannot be obtained in one step.

The EM algorithm can be employed to calculate the maximum likelihood estimates (MLE) of the isoform proportions $P = (p_1, p_2, \dots, p_j)$ from our observed cDNA fragment-compatible matrix Y. The EM algorithm works in an iterative way, and it will be converged after

numbers of iterations. Let's use $P^{(k)}$ to denote the isoform proportions computed after kth iteration. We initialized $P^{(0)} = (\, p_j{}^{(0)}, j = 1, \dots, J\,)$ as $p_j{}^{(0)} = 1/J$. Each iteration updates $P^{(k)}$ to $P^{(k+1)}$ through accomplishing the following E and M steps:

E-step:

$$z_{ij}{}^{(k+1)} = E\,[\, z_{ij} \mid Y_i, P^{(k)}\,] = Pr\,(\, z_{ij} = 1 \mid Y_i, P^{(k)})$$

$$= z_{i,j}^{(k+1)} = \frac{y_{i,j} \times p_j^{(k)}}{\sum_{j=1}^{J} y_{i,j} \times p_j^{(k)}}, \forall i, j.$$

M-step:

Let $n_j^{(k+1)} = \sum_{i=1}^{N} z_{i,j}^{(k+1)}, \forall j,$

$$p_j^{(k+1)} = \frac{n_j^{(k+1)}}{N}, \forall j.$$

The E-step updates the probabilities that each short read generated from isoform I_j based on the current estimated isoform proportion set $P^{(k)}$ from $z_{ij}{}^{(k)}$ to $z_{ij}{}^{(k+1)}$, and M-step updates isoform proportion set from $P^{(k)}$ to $P^{(k+1)}$ based on $z_{ij}{}^{(k+1)}$. The EM algorithm iterates between E and M steps until convergence, i.e.,$\sum_{j=1}^{J} |p_j^{(k+1)} - p_j^{(k)}| < \epsilon$, ϵ is an arbitrarily small positive number, i.e. 0.00001. To this end, we get the converged isoform proportions as $P^{(k+1)} = (\, p_j{}^{(k+1)}, j = 1, \dots, J\,)$. The EM algorithm for isoform quantification using RNA-seq is adapted from a similar algorithm developed for isoform quantification using EST data (Xing et al 2006).

In order to test the accuracy of our EM algorithm in estimating isoform proportions, we used simulation data generated from FluxSimulator. The details are described in our manuscript.

Section 2. Additional evidence for superiority of the seed enrichment approach over the seed presence approach

The seed enrichment approach predicts targets from a statistical perspective by comparing "seed concentration" in different genomic regions for each isoform with "expected" seed concentration in the whole genome range. We hypothesize that the higher seed concentration in a particular region, the higher chance one of the seeds will be functional in that region. In order to test this hypothesis, we provide a set of complementary evidence as follows:

(1). Comparing with the seed presence approach, our seed enrichment approach represents a more stringent way to miRNA target prediction. As demonstrated in Figure S3, the seed enrichment approach further reduces the number of candidate targets compared with the seed presence approach. It helps researchers, especially experimentalists, focus on the highest seed concentrated region and provides biological insights on miRNA-155 binding mechanisms.

Moreover, we found that the result of our seed enrichment analysis is consistent with the existing biological evidence (e.g. Yue D et al., *Curr Genomics.*, 10(7):478-92, 2009.) in that seeds in the 3'-UTR are more potent than those in the 5'-UTR or coding region.

To summarize, using the seed enrichment approach, we are not attempting to predict which seed is the functional one. Instead we attempt to predict a specific seed site region, where is more likely to contain a functional seed.

(2). To compare seed presence approach and seed enrichment approach in an objective way, we examined the performance of both approaches using a published dataset reported in *Bandyopadhyay and Mitra (Bioinformatics, 25(20):2625-2631, 2009*). In the dataset, we used human miRNA targets having seed(s) present in their 3'-UTR's: 106 of these are experimentally verified positive targets, and 10 isoforms are experimentally verified negative targets. If we were to use the seed presence approach, all of these 116 isoforms would have been predicted as positive targets. In comparison, after using seed enrichment approach, we are able to discriminate those 116 targets into both positive and negative groups. The positive group enrichment p-value is 0.000676, and the negative group enrichment p-value is 0.996867 (Pearson's Chi-Square Test). The p-values indicate that positive targets tend to have seed enrichment in their 3'-UTR's, while negative targets do not. In summary, the seed enrichment approach is more accurate and specific than the seed presence approach in predicting microRNA targets. We also include this analysis to the supplemental material of this manuscript (Table S7).

Section 3: Additional justification of combining seed enrichment and isoform-level down-regulation in predicting miRNA targets.

We provide justification sequentially as follows. **First,** we consider the situation where all the miRNA-155 targets, either predicted by the isoform-level computational approach or by the seed enrichment/presence approach, are in the 3'-UTR's. Our goal is to validate our isoform-level approach combining down-regulation and seed enrichment. **Second,** we extend our miRNA-155 target prediction to other regions of the transcripts. Our goal is two-fold: to confirm 3'-UTR is still the primary targeting region for miRNA-155; to predict other regions as potential targeting regions as well, which is consistent to the recent findings (e.g. *Lytle et al., PNAS, 104(23): 9667-72, 2007; Inhan Lee et al., Genome Res., 19(7):1175-83, 2009* and *Yue D et al., Curr Genomics., 10(7):478-92, 2009*).

Targeting 3'-UTR region: As we mentioned in the manuscript, transcript down-regulation along does not provide sufficient condition to be considered as a transcript targeting criterion due to indirect cellular responses or other binding mechanisms. Similarly, because some seeds are non-functional, we cannot avoid false targets if we only use seed enrichment as the sole prediction criterion. It follows that neither of these two criteria are good enough for miRNA-155

target prediction at isoform-level. Based on the conventional notion that if one seed is functional, it must have a site in the 3'-UTR of the target transcript, causing the transcript down regulated. Thus, we believe that combining differential expression analysis at isoform-level and seed enrichment analysis will lead to more accurate target prediction and further reduce the number of candidate targets. The stringent criteria of combining down-regulation and seed enrichment are particularly meaningful to experimentalists.

Based on the reviewer's comment, we investigated the number of target transcripts under different criteria with the down-regulation ratio set to 0.8. The result is shown in Figure S3:

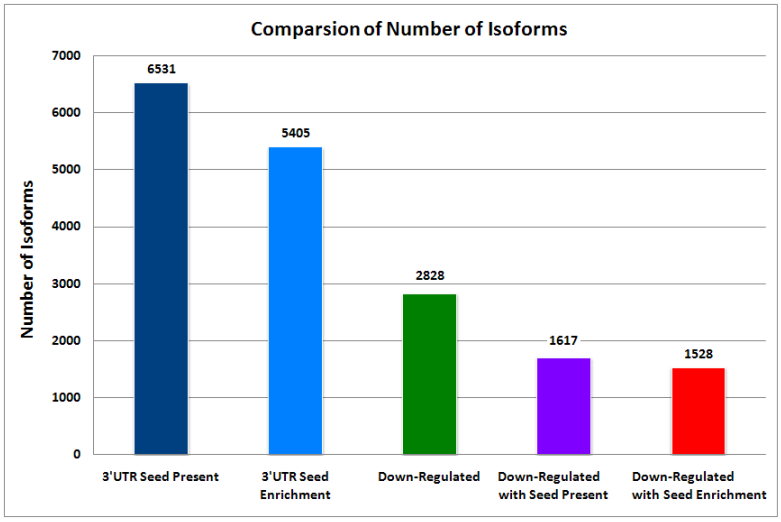

Figure S3. Number of isoform transcripts satisfying different criteria, from the most relaxed to the most stringent.

In Figure S3, we can clearly see the advantage of combining differential expression analysis and seed enrichment to predict miR-155 targets at isoform-level. There are 6,531 isoforms with at least one seed presenting at their 3'-UTR's. By using the seed enrichment approach, we can further decrease this number to 5,405. 2,828 isoforms show down-regulation by miR-155. As shown in Figure S3, by using a combination of down-regulation with 3'-UTR seed enrichment, we are able to decrease the number of predicted isoform targets down to 1,528.

Targeting 3'-UTR, 5'-UTR and coding region: Since it has been recently reported that miRNA binds not only to 3'-UTR but also to 5'-UTR and the coding region (e.g. *Lytle et al., PNAS, 104(23):9667-72, 2007; Inhan Lee et al., Genome Res., 19(7):1175-83, 2009.* and *Yue D et al., Curr Genomics., 10(7):478-92, 2009.*), we also performed a genome-wide seed enrichment analysis. The percentage of seed enrichment regions are shown in Figure S4.

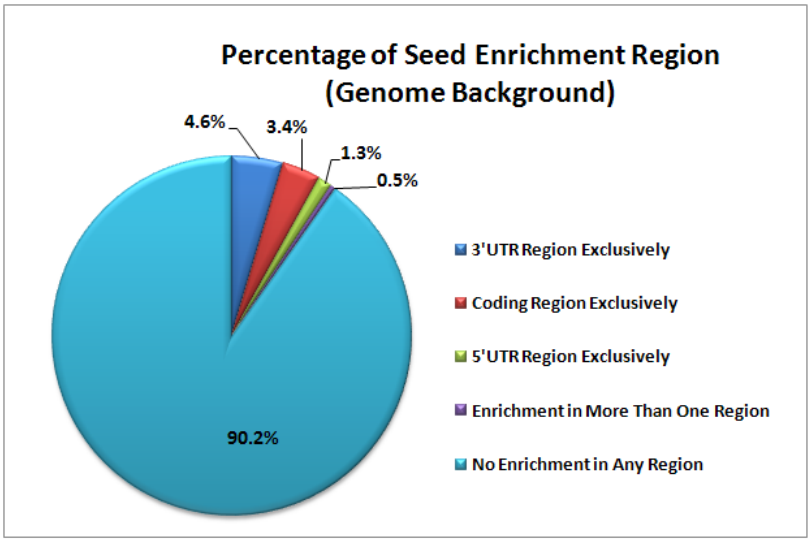

Figure S4. miRNA-155 seed enrichment regions using the genome background

From Figure S4, there are only about 10% of the total numbers of isoforms having seed enrichment in 3'-UTR, 5'-UTR or coding region. The isoforms with seed enrichment in 3'-UTR region exclusively (4.6%) have a higher percentage than those in 5'-UTR (1.3%) or coding region (3.4%) exclusively. The isoforms with seed enrichment in more than one region are very rare (0.5%). Figures S3, S4 and additional analysis illustrate that the seed enrichment criterion is expected to be highly accurate and specific.

Section 4: The EM algorithm fails to accurately estimate abundance levels of around 10% of isoforms.

We simulated both single-ended and paired-end short reads using FluxSimulator, and used EM algorithm to estimate the abundance of isoforms in each gene (Figure 2b for single-ended short reads, and Figure S1 for paired-end short reads). By using simulation data, both figures show that our EM algorithm can estimate isoform abundance very accurately since the estimated abundances are highly correlated with the true abundances for the vast majority of the isoforms with high robust R^2. The very small portion of purple dots moving further away from the regression line (red line) correspond to those isoforms for which the EM algorithm fails to estimate their abundance accurately in some situations, such as too many isoforms within one gene, the length of the unique exon is shorter than the read length, and so on.

We investigated the isoforms whose abundances were not accurately estimated by EM algorithm. The isoforms that meet the requirement (|log10(trueRPKM) – log10(predictedRPKM)| >= 1, RPKM is a standard way to represent the abundance of gene expression) are considered as outliers. There are about 10% of isoforms falling into this category. We further examined these outliers. We found that nearly 74% of outlier isoforms belong to genes with more than 5 isoforms and about 41% of outlier isoforms have at least one exon whose

length is less than 50 bases. This analysis clearly demonstrates a limitation of our EM algorithm in estimating isoform abundance; however, it is predominantly accurate and effective for around 90% of isoforms in the transcriptome.

Section 5: Discovery of novel transcripts.

First we performed exon junction analysis using Tophat (23), and we detected 2,553 novel junctions that are not annotated in ASTD database. 544 out of 2,553 novel junctions share exons with the annotated transcripts through skipping exon or mutually exclusive exons mechanisms, corresponding to a total of 499 genes (Both annotated and novel transcripts relevant to these 499 genes are included in Table S5). We verified 9 out of 10 selected novel junctions using quantitative RT-PCR experiments (Table II, Figure S5). Second, we composed 1,572 novel transcripts that are supported by these 544 novel junctions as illustrated in Figure 1b. We augmented the ASTD annotation table by merging the 1,572 novel transcripts, followed by isoform quantification analysis using EM algorithm. After filtering by isoform RPKM value at 0.2 cutoff, we discovered 51 significantly down-regulated novel transcripts with seed enrichment at 3'-UTR region, which are highly likely to be microRNA-155 targets. Detailed results are included in Table S6. We gave an example novel transcript in Figure S2.

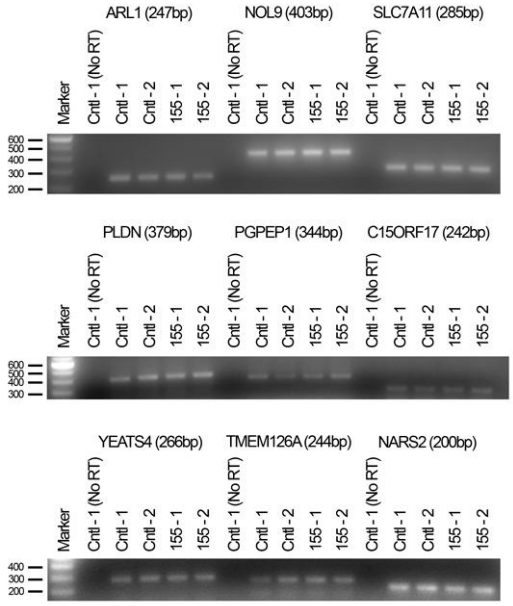

Figure S5. Verification of selected novel junctions using quantitative RT-PCR experiments. 9 out of 10 were verified.

Section 6: Legends of all supplemental tables.

Table S1. Lists of putative microRNA-155 targets falling into each of the eight categories using down-regulation cut-off values of 0.6, 0.7 and 0.8.

Table S2. A list of predicted microRNA-155 targets through whole transcriptome studies (down-regulation cut-off of 0.8).

Table S3. A list of predicted microRNA-155 targets through whole transcriptome studies (down-regulation cut-off of 0.7).

Table S4. A list of predicted microRNA-155 targets through whole transcriptome studies (down-regulation cut-off of 0.6).

Table S5. A list of annotated and novel transcripts supported by the context-specific junction evidence.

Table S6. A list of predicted novel (previously not annotated) microRNA-155 targets.

Table S7. Comparing the seed presence approach and the seed enrichment approach in an objective way using a published dataset reported in *Bandyopadhyay and Mitra (Bioinformatics, 25(20):2625-2631, 2009*)

References

Bandyopadhyay S and Mitra R. (2009) TargetMiner: microRNA target prediction with systematic identification of tissue-specific negative examples. *Bioinformatics,* 25(20):2625-31.

Lee I, Ajay SS, Yook JI, Kim HS, Hong SH, Kim NH, Dhanasekaran SM, Chinnaiyan AM and Athey BD. (2009) New class of microRNA targets containing simultaneous 5'-UTR and 3'-UTR interaction sites. *Genome Research*, 19(7): 1175-83.

Lytle JR, Yario TA and Steitz JA. (2007) Target mRNAs are repressed as efficiently by microRNA-binding sites in the 5' UTR as in the 3' UTR. *PNAS*, 104(23):9667-72.

Yue D, Liu H and Huang Y. (2009) Survey of computational algorithms for microRNA target prediction. *Current Genomics*, 10(7): 478–92.

Xing Y, Yu T, Wu NY, Roy M, Kim J and Lee C. (2006) An expectation-maximization algorithm for probabilistic reconstruction of full-length isoforms from splice graphs. Nucleic Acids Research, 34, 3150-3160.